Jambura Journal of Mathematics
Volume 1 Nomor 2, Juli 2019
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jjom , P-ISSN: 2654-5616, E-ISSN: 2656-1344

Penerapan *Hierarchical Clustering* Metode *Agglomerative* pada Data Runtun Waktu

Andrea Tri Rian Dani[1*], Sri Wahyuningsih[2], Nanda Arista Rizki[3]

[1,2] *Program Studi Statistika, Jurusan Matematika, Fakultas MIPA, Universitas Mulawarman, Jl. Barong Tongkok No. 04 Gunung Kelua, Kota Samarinda 75119, Kalimantan Timur, Indonesia*

[3] *Program Studi Matematika, Jurusan Matematika, Fakultas MIPA Universitas Mulawarman, Jl. Barong Tongkok No. 04 Gunung Kelua, Kota Samarinda 75119, Kalimantan Timur, Indonesia*

* *Penulis Korespondensi. Email: andrikadoko@gmail.com*

ABSTRAK

Analisis *cluster* merupakan seperangkat metode yang digunakan untuk mengelompokkan objek ke dalam sebuah *cluster* berdasarkan informasi yang ditemukan pada data. Analisis *cluster* dapat diterapkan pada data runtun waktu, di mana terdapat prosedur dan algoritma pengelompokkan yang berbeda dibandingkan dengan pengelompokkan data *cross-section*. Banyak teknik pengelompokkan data runtun waktu yang dikembangkan di antaranya adalah penggunaan jarak pengukuran kemiripan yang sesuai dengan karakteristik data runtun waktu, pemilihan algoritma pengelompokkan yang optimal sampai dengan penentuan banyaknya *cluster* yang representatif. Tujuan dari penelitian adalah untuk memperoleh jarak pengukuran kemiripan terbaik, kemudian memperoleh algoritma pengelompokkan metode *agglomerative* yang optimal serta memperoleh jumlah *cluster* yang representatif. Pemilihan jarak pengukuran kemiripan terbaik dan algoritma yang optimal menggunakan koefisien korelasi *cophenetic*, sedangkan untuk penentuan jumlah *cluster* menggunakan koefisien *silhouette*. Data pada penelitian adalah data jumlah penduduk Kabupaten/Kota di Provinsi Kalimantan Timur dari Tahun 2005-2017. Berdasarkan hasil analisis, diperoleh jarak pengukuran kemiripan terbaik dalam mengelompokkan Kabupaten/Kota di Provinsi Kalimantan Timur adalah jarak *autocorrelation based distance* (ACF) dengan nilai koefisien korelasi *cophenetic* sebesar 0,99. Algoritma pengelompokkan yang optimal adalah algoritma *average linkage*, dikarenakan memiliki nilai koefisien korelasi *cophenetic* yang terbesar diantara algoritma pengelompokkan lainnya, dengan jumlah *cluster* yang representatif berdasarkan koefisien *silhouette* adalah 2 *cluster*.

Kata Kunci:
Autocorrelation Based Distance (ACF); *Hierarchical Clustering*; Koefisien Korelasi *Cophenetic*; Koefisien *Silhouette*

Diterima:	*Disetujui:*	*Online:*
01-06-2019	*28-06-2019*	*30-07-2019*

ABSTRACT

Cluster analysis is a set of methods used to groups objects into a cluster based on information from data. Cluster analysis can be applied to time series data by performing several different procedures and algorithms than grouping cross-section data. Many techniques are developed along with research on clustering of time series data, including selection of similarity distances that are adjusted to the characteristics of the time series, selection of optimal grouping algorithms, and determination for the number of representative clusters. The purpose of this study was to obtain the best distance measurement similarity, to obtain an optimal grouping algorithm based on agglomerative methods, and to obtain a representative number of clusters. The best similarity distance and optimal algorithm were selected using coefficient of cophenetic correlation, while the number of clusters was determined using the silhouette

coefficient. The data used in this study was the population of regencies or cities in East Kalimantan province from 2005 to 2017. Based on the results of analysis, the best similarity distance was the autocorrelation based distance (ACF) where the coefficient of cophenetic correlation was 0.99. The optimal grouping algorithm was the average linkage algorithm, because it had the largest coefficient of cophenetic correlation among coefficients for other grouping algorithms. The representative number of clusters based on the silhouette coefficient were 2 clusters.

Keywords:
Autocorrelation Based Distance (ACF); Hierarchical Clustering; Cophenetic Correlation Coefficient; Silhouette Coefficient

Received:	***Accepted:***	***Online:***
2019-06-01	*2019-06-28*	*2019-07-30*

DOI: http://dx.doi.org/10.34312%2Fjjom.v1i2.2354

1. Pendahuluan

Seiring dengan majunya perkembangan teknologi informasi saat ini, *data mining* merupakan salah satu ilmu yang berkembang sangat pesat, dikarenakan besarnya kebutuhan akan informasi lebih dari sekumpulan data berskala besar. Pertumbuhan yang sangat pesat dari sekumpulan data telah menciptakan suatu kondisi dimana kaya akan data tetapi informasi yang dimiliki sangat minim. *Data mining* dapat digunakan untuk menggali suatu penemuan baru dengan mencari struktur atau pola tertentu dari sekumpulan data berskala besar [1]. Pekerjaan yang berkaitan dengan *data mining* antara lain: pemodelan prediktif, analisis *cluster*, analisis asosiasi dan deteksi anomali [2].

Analisis *cluster* merupakan salah satu alat yang penting dalam pengolahan data statistik untuk melakukan analisis data. Analisis *cluster* merupakan seperangkat metode yang digunakan untuk mengelompokkan objek ke dalam sebuah *cluster* berdasarkan informasi yang ditemukan pada data. Hasil *cluster* dikatakan baik ketika mempunyai homogenitas yang besar antar objek dalam satu *cluster* dan heterogenitas yang besar pula antar *cluster* yang satu dengan *cluster* lainnya [2].

Analisis *cluster* dapat diaplikasikan pada data runtun waktu, di mana terdapat prosedur dan algoritma pengelompokan yang berbeda dibandingkan dengan pengelompokan data *cross section*. Algoritma dan prosedur dalam proses pembentukan *cluster* dilakukan berbeda, karena data runtun waktu adalah suatu kumpulan data yang terjadi berdasarkan waktu secara runtun (terurut) dengan interval waktu yang konstan. Kumpulan data runtun waktu dapat dikelompokkan berdasarkan karakteristik pada masing-masing data runtun waktu tersebut menggunakan analisis *cluster* runtun waktu, dengan cara mengelompokkan objek berdasarkan pola runtun waktunya [3]. Saat ini analisis *cluster* telah banyak digunakan di berbagai bidang ilmu pengetahuan seperti ekonomi, psikologi, kesehatan, sosial masyarakat dan kependudukan.

Provinsi Kalimantan Timur (Kaltim) merupakan Provinsi terluas kedua setelah Papua di Indonesia. Provinsi Kaltim memiliki potensi sumberdaya alam yang besar, akan tetapi sebagian besar potensi tersebut belum digunakan secara maksimal. Salah satu faktor dalam masalah kemiskinan yaitu jumlah penduduk yang tinggi. Peningkatan jumlah penduduk tanpa diiringi dengan kemajuan faktor-faktor lainnya tentu akan memunculkan masalah baru, salah satunya adalah minimnya lapangan pekerjaan dan tingkat kemakmuran yang rendah. Dalam arti sederhana, jumlah penduduk adalah banyaknya manusia yang bertempat tinggal atau berdomisili pada suatu wilayah atau

daerah dan memiliki pekerjaan yang tetap serta terdaftar secara sah berdasarkan peraturan perundang-undangan yang berlaku.

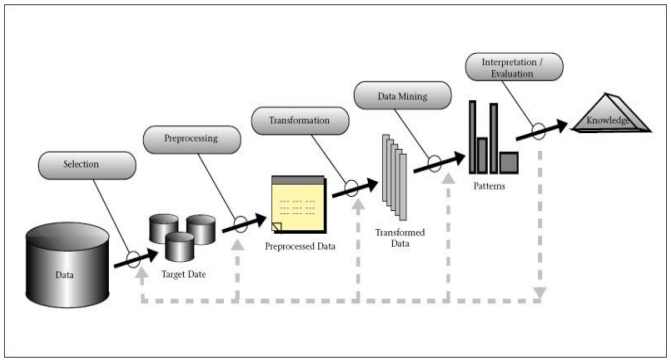
Berdasarkan latar belakang di atas, penulis tertarik untuk membahas mengenai analisis *cluster* pada proses pengelompokkan Kabupaten/Kota Provinsi Kaltim berdasarkan jumlah penduduk dengan data runtun waktu. Pada penelitian ini, digunakan 2 jarak pengukuran kemiripan yaitu *autocorrelation based distance* (ACF) dan *dynamic time warping* (DTW). Metode pengelompokkan yang digunakan adalah pengelompokkan hierarki dengan semua algoritma yang ada pada metode *agglomerative* (pemusatan). Tujuan dari penelitian ini adalah memperoleh hasil pengelompokkan yang optimal pada proses pengelompokkan Kabupaten/Kota Provinsi Kaltim berdasarkan jumlah penduduk dengan data runtun waktu.

2. Metode

2.1. *Data Mining*

Data mining adalah sekumpulan metode untuk menemukan pola yang tersirat dari suatu kumpulan data sehingga nantinya diperoleh informasi yang bermanfaat. *Data mining* juga merupakan proses mengekstraksi informasi yang menarik, implisit dan berpeluang untuk dimanfaatkan dari data berskala besar (*big data*). Jenis pembelajaran yang terdapat dalam *data mining* yaitu *supervised learning* dan *unsupervised learning*. Pembelajaran *supervised learning* digunakan untuk memperkirakan suatu nilai, sedangkan pembelajaran *unsupervised learning* digunakan untuk mencari suatu struktur intrinsik, hubungan dalam suatu data yang tidak membutuhkan *class* sebelum dilakukan proses pembelajaran [4].

Tahapan-tahapan dalam proses *data mining* ditampilkan pada Gambar 1 [5].



Gambar 1. Tahapan proses *data mining*

2.2. Analisis *Cluster*

Analisis *cluster* adalah salah satu alat bantu pada proses *data mining* yang bertujuan untuk mengelompokkan objek-objek ke dalam suatu *cluster*. *Cluster* itu sendiri adalah sekelompok atau sekumpulan objek data yang memiliki kemiripan satu sama lain dalam kelompok yang sama. Tujuan analisis *cluster* adalah tidak untuk mengkorelasikan objek yang satu dengan objek lainnya, melainkan untuk mengidentifikasi sekelompok objek yang mempunyai kesamaan dan karakteristik khas yang dapat dipisahkan dengan kelompok lainnya. Objek yang berada pada *cluster* yang sama relatif lebih homogen daripada objek yang berada pada *cluster* yang berbeda. Jumlah kelompok yang dapat diidentifikasi tergantung pada banyak dan variansi data objek [2].

Hasil *cluster* dikatakan baik dengan ciri-ciri sebagai berikut:
1. Memiliki homogenitas yang besar antar anggota pada *cluster* yang sama (*within-cluster*).
2. Memiliki heterogenitas yang besar antar *cluster* yang satu dengan *cluster* lainnya (*between-cluster*) [6].

2.3. Prosedur Analisis *Cluster*

Adapun tahapan-tahapan dalam prosedur analisis *cluster* sebagai berikut:

1. Merumuskan masalah

Urgensi dalam melakukan perumusan masalah analisis *cluster* adalah pemilihan variabel yang nantinya digunakan dalam proses pengelompokkan (pembentukan *cluster*). Pada dasarnya variabel-variabel yang akan dipilih harus menguraikan kemiripan (*similarity*) antar objek dan benar-benar relevan dengan masalah yang dibahas. Variabel yang dipilih harus berdasarkan suatu pertimbangan yang berkenaan dengan dugaan yang akan diteliti [7].

2. Normalisasi data

Tujuan dari proses analisis *cluster* adalah untuk mengelompokkan objek-objek yang mempunyai kemiripan yang sama dalam satu *cluster*. Objek dengan jarak yang lebih dekat akan lebih mirip satu sama lain dibandingkan jarak yang lebih jauh. Jika rentang nilai antar objek memiliki perbedaan skala yang cukup besar yang dapat menyebabkan bias dalam analisis *cluster*, maka data asli perlu dilakukan normalisasi. Normalisasi dapat menyingkirkan atau menghilangkan pengaruh dari unit pengukuran dan dapat memperkecil perbedaan antar kelompok atau *cluster* [7].

Normalisasi data dapat dilakukan dengan cara semua dimensi atau sub-variabel penyusun ditransformasi ke dalam data standar (nilai rata-rata sama dengan nol, variansi sama dengan satu). Cara menentukan nilai normalisasi adalah dengan menghitung nilai rata-rata dan deviasi standar yaitu:

$$\bar{Z} = \left(\frac{1}{n}\right)\sum_{t=1}^{n} Z(t) \tag{1}$$

dan

$$S_Z = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(Z(t) - \bar{Z})^2} \tag{2}$$

kemudian menghitung data hasil normalisasi dengan menggunakan Persamaan (3)

$$\tilde{Z}(t) = \frac{Z(t) - \bar{Z}}{S_Z} \tag{3}$$

dengan,
$Z(t)$: data Z pada waktu ke-t
n : banyaknya data
$\bar{Z}$: rata-rata dari $Z(t)$
S_Z : deviasi standar dari $Z(t)$
$\tilde{Z}(t)$: normalisasi data Z pada waktu ke-t [8].

2.4. Pengukuran Kemiripan

Pada dasarnya proses pembentukan *cluster* yaitu mencari dan mengelompokkan objek-objek berdasarkan kemiripan dan kedekatan antar objek yang satu dengan objek lainnya. Langkah awal yaitu mengukur seberapa dekat kemiripan dan kedekatan antar objek tersebut. Adapun pengukuran kemiripan yang digunakan untuk mengelompokkan data runtun waktu pada penelitian ini adalah:

1. *Dynamic time warping* (DTW)

Dynamic time warping adalah algoritma untuk mencari *warping path* yang optimal antara dua data runtun waktu sehingga *output*-nya adalah kumpulan nilai-nilai *warping path* serta jarak di antara kedua data runtun waktu tersebut. Algoritma *dynamic time warping* dapat digunakan untuk mengukur kedekatan dua data runtun waktu dengan jumlah data yang berbeda. *Dynamic time warping* menggunakan teknik pemrograman dinamis untuk menemukan semua jalur yang mungkin dan memilih salah satu yang menghasilkan jarak minimum antara dua data runtun waktu menggunakan matriks jarak.

Misalkan terdapat dua data runtun waktu dengan panjang yang berbeda yaitu $Z(t) = Z(1), Z(2), \ldots, Z(i), \ldots, Z(m)$ dengan $Y(t) = Y(1), Y(2), \ldots, Y(j), \ldots, Y(n)$. Langkah awal adalah membuat matriks **C** berukuran $n \times m$. Elemen ke-(i,j) dalam matriks **C** didefinisikan sebagai selisih antara $Z(i)$ dengan $Y(j)$, kemudian ditambah dengan nilai minimum tiga elemen yang berdekatan $\{c_{(i-1)(j-1)}, c_{(i-1),j}, c_{(i)(j-1)}\}$, di mana $0 < i \leq m$ dan $0 < j \leq n$. Elemen ke-(i,j) dalam matriks **C** dapat ditulis menjadi

$$c_{ij} = w_{ij} + min\{c_{(i-1)(j-1)}, c_{(i-1),j}, c_{(i)(j-1)}\} \tag{4}$$

Dalam hal ini nilai w_{ij} merupakan selisih antara $Z(i)$ terhadap $Y(j)$ dengan perhitungan dapat dituliskan pada Persamaan (5).

$$w_{ij} = |Z(i) - Y(j)| \tag{5}$$

Berdasarkan Persamaan (4) dan (5), maka jarak *dynamic time warping* antara dua data runtun waktu $Z(t)$ terhadap $Y(t)$ dapat didefinisikan

$$d_{DTW}(Z,Y) = \min_{\forall w \in P} \left\{ \sqrt{\sum_{i,j=1}^{K} c_{ij}} \right\} \tag{6}$$

dengan *P* adalah sekumpulan dari semua *warping path* yang mungkin, c_{ij} adalah elemen (i,j) pada *warping path* serta *K* adalah panjang dari *warping path* [9].

2. *Autocorrelatioan based distance*

Galeano dan Pena (2000) melakukan penelitian mengenai hubungan data runtun waktu dengan menggunakan pendekatan fungsi otokorelasi (FOK). Ilustrasi untuk perhitungan jarak FOK adalah sebagai berikut, misalkan diberikan dua data runtun waktu dengan ukuran n yaitu $Z(t) = Z(1), Z(2), \ldots, Z(n)$ dan $Y(t) = Y(1), Y(2), \ldots, Y(n)$ sehingga $\hat{\boldsymbol{\rho}}_{\boldsymbol{Z}} = (\hat{\rho}_Z(1), \hat{\rho}_Z(2), \ldots, \hat{\rho}_Z(n))'$ dan $\hat{\boldsymbol{\rho}}_{\boldsymbol{Y}} = (\hat{\rho}_Y(1), \hat{\rho}_Y(2), \ldots, \hat{\rho}_Y(n))'$ adalah vektor-vektor otokorelasi hasil pendugaan dari data runtun waktu $Z(t)$ dan $Y(t)$. Jarak antara dua runtun waktu kemudian dapat dibentuk pada Persamaan (7) sebagai berikut:

$$d_{FOK}(Z,Y) = \sqrt{(\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}} - \widehat{\boldsymbol{\rho}}_{\boldsymbol{Y}})'\boldsymbol{\Omega}(\widehat{\boldsymbol{\rho}}_{\boldsymbol{Z}} - \widehat{\boldsymbol{\rho}}_{\boldsymbol{Y}})} \quad (7)$$

dengan $d_{FOK}(Z,Y)$ adalah jarak otokorelasi antara data runtun waktu $Z(t)$ terhadap $Y(t)$, sedangkan $\boldsymbol{\Omega}$ adalah matriks identitas [10].

2.5. Metode dalam Analisis *Cluster*

Pada analisis *cluster* terdapat dua metode yang dapat digunakan, yaitu metode hierarki dan non-hierarki. Pengelompokkan hierarki (*hierarchical clustering*) adalah metode analisis *cluster* dengan cara membangun sebuah hierarki kelompok. Strategi untuk pengelompokkan hierarki pada umumnya dibagi menjadi dua jenis yaitu *agglomerative* (pemusatan) dan *divisive* (penyebaran). Metode *agglomerative* (pemusatan) merupakan metode pengelompokkan hierarki dengan pendekatan bawah-atas (*bottom up*). Metode *agglomerative* (pemusatan) biasanya dipergunakan pada bidang ekonomi dan sosial masyarakat [11]. Adapun algoritma-algoritma pada pengelompokkan hierarki dengan metode *agglomerative* (pemusatan) sebagai berikut:

1. *Single linkage*

Pautan tunggal (*single linkage*) merupakan prosedur pengelompokkan *agglomerative* berdasarkan jarak terkecil antar objek. Algoritma pengelompokan *single linkage* diawali dengan memilih jarak terkecil dalam matriks $\boldsymbol{D} = \{d_{ij}\}$, kemudian menggabungkan objek yang bersesuaian misalnya U dan V untuk memperoleh *cluster* (UV). Langkah berikutnya adalah mencari jarak antara (UV) dengan *cluster* lainnya, misalnya W sehingga dapat dituliskan sebagai berikut:

$$d_{(UV)W} = min(d_{UW}, d_{VW}) \quad (8)$$

dengan d_{UW} adalah jarak tetangga terdekat dari *cluster* U dan W serta d_{VW} adalah jarak tetangga terdekat dari *cluster* V dan W [11].

2. *Complete linkage*

Pautan penuh (*complete linkage*) merupakan prosedur pengelompokan *agglomerative* berdasarkan jarak terbesar antar objek. Algoritma *complete linkage* diawali dengan memilih jarak terbesar dalam matriks $\boldsymbol{D} = \{d_{ij}\}$, kemudian menggabungkan objek yang bersesuaian misalnya U dan V untuk memperoleh *cluster* (UV). Langkah berikutnya adalah mencari jarak antara (UV) dengan *cluster* lainnya, misalnya W sehingga dapat dituliskan sebagai berikut:

$$d_{(UV)W} = max(d_{UW}, d_{VW}) \quad (9)$$

dengan d_{UW} adalah jarak tetangga terjauh dari *cluster* U dan W serta d_{VW}adalah jarak tetangga terjauh dari *cluster* V dan W [11].

3. *Average linkage*

Pautan rata-rata (*average linkage*) merupakan prosedur pengelompokan *agglomerative* berdasarkan rata-rata antar objek. Algoritma *average linkage* diawali dengan mendefinisikan matriks $\boldsymbol{D} = \{d_{ij}\}$ untuk memperoleh objek yang paling dekat, sebagai contoh U dan V, kemudian objek ini digabung ke dalam bentuk *cluster* (UV) dan selanjutnya jarak antara (UV) dengan *cluster* lainnya W, sehingga dapat dituliskan sebagai berikut:

$$d_{(UV)W} = \frac{d_{(UW)} + d_{(VW)}}{n_{(UV)} n_W} \tag{10}$$

dengan $n_{(UV)}$ adalah banyaknya anggota dalam *cluster* (UV) dan n_W adalah banyaknya anggota dalam *cluster* W [11].

4. *Ward method*

Ward method berusaha untuk meminimumkan variasi antar objek yang berada pada satu *cluster*. Algoritma *ward* dimulai dengan mendefinisikan matriks $\boldsymbol{D} = \{d_{ij}\}$ untuk memperoleh objek yang paling mirip, sebagai contoh U dan V, kemudian objek ini digabung ke dalam bentuk *cluster* (UV) dan selanjutnya jarak antara (UV) dengan *cluster* lainnya W, sehingga dapat dituliskan sebagai berikut:

$$d_{(UV)W} = \frac{\left[(n_W + n_U) d_{(UW)} + (n_W + n_V) d_{(VW)}\right] - n_W d_{(UV)}}{n_W + n_{(UV)}} \tag{11}$$

Langkah pembentukan *cluster* akan terus berulang, demikian seterusnya sampai semua objek bergabung dalam jumlah *cluster* yang ditentukan [12].

5. *WPGMA method*

Mcquitty method atau biasa disebut *Weighted Pair Group Method with Arithmetic Mean* (WPGMA) adalah salah satu algoritma yang dapat digunakan dalam proses pengelompokkan hierarki metode *agglomerative* (pemusatan). Algoritma WPGMA dimulai dengan mendefinisikan matriks $\boldsymbol{D} = \{d_{ij}\}$ untuk memperoleh objek yang paling mirip, sebagai contoh U dan V, kemudian objek ini digabung ke dalam bentuk *cluster* (UV) dan selanjutnya jarak antara (UV) dengan *cluster* lainnya W, sehingga dapat dituliskan sebagai berikut:

$$d_{(UV)W} = \frac{d_{(UW)} + d_{(VW)}}{2} \tag{12}$$

Langkah pembentukan *cluster* akan terus berulang, demikian seterusnya sampai semua objek bergabung dalam jumlah *cluster* yang ditentukan [12].

6. *Median method*

Median method merupakan metode pengelompokkan dengan memperhatikan median dari setiap objek yang bergabung berdasarkan jarak minimum yang diperoleh dari matriks $\boldsymbol{D} = \{d_{ij}\}$. Algoritma *median method* hampir sama dengan algoritma *centroid method*, hanya saja yang membedakan adalah perlu menghitung median dari *cluster* U dan V berdasarkan Persamaan 13.

$$d_{(UV)W} = \frac{d_{(UW)} + d_{(VW)}}{2} - \frac{d_{(UV)}}{4} \tag{13}$$

Langkah pembentukan *cluster* akan terus berulang, demikian seterusnya sampai semua objek bergabung dalam jumlah *cluster* yang ditentukan [12].

7. *Centroid method*

Centroid method adalah rata-rata semua objek dalam *cluster*. Jarak antar kedua *cluster* adalah jarak antar *centroid cluster* itu sendiri. *Centroid cluster* merupakan nilai tengah

observasi pada suatu variabel dalam suatu *set*. Algoritma *centroid* dimulai dengan mendefinisikan matriks $\boldsymbol{D} = \{d_{ij}\}$ untuk memperoleh jarak objek yang paling mirip, sebagai contoh U dan V. *Centroid cluster* yang baru terbentuk berdasarkan Persamaan 14.

$$d_{(UV)W} = \frac{n_U d_{(UW)} + n_V d_{(VW)}}{n_{(UV)}} - \frac{n_U n_V d_{(UV)}}{{n_{(UV)}}^2} \tag{14}$$

Pada saat objek digabungkan maka *centroid* baru dihitung, sehingga setiap kali ada penambahan anggota *centroid* akan berubah pula [12].

2.6. Validitas *Cluster*

Adapun uji validitas yang digunakan dalam penelitian ini sebagai berikut:

1. Koefisien korelasi *cophenetic*

Koefisien korelasi *cophenetic* merupakan koefisien korelasi antara elemen-elemen asli matriks ketidakmiripan (*dissimilarity distance*) dan elemen-elemen yang dihasilkan oleh dendogram (matriks *cophenetic*). Koefisien korelasi *cophenetic* dapat dihitung berdasarkan Persamaan 15.

$$r_{coph} = \frac{\sum_{i<j}^{n}(d_{ij} - \bar{d})(d_{coph\sim ij} - \bar{d}_{coph})}{\sqrt{\left[\sum_{i<j}^{n}(d_{ij} - \bar{d})^2\right]\left[\sum_{i<j}^{n}(d_{coph\sim ij} - \bar{d}_{coph})^2\right]}} \tag{15}$$

dengan,

r_{coph} : koefisien korelasi *cophenetic*
d_{ij} : jarak asli antara objek ke-*i* dan ke-*j*
$\bar{d}$: rata-rata d_{ij}
$d_{coph\sim ij}$: jarak *cophenetic* objek ke-*i* dan ke-*j*
$\bar{d}_{coph}$: rata-rata $d_{coph\sim ij}$.

Nilai r_{coph} berkisar antara -1 dan 1, nilai r_{coph} mendekati 1 berarti proses *clustering* yang dihasilkan dapat dikatakan cukup baik [13].

2. Koefisien *silhouette*

Salah satu metode evaluasi yang digunakan untuk melihat optimalisasi dari suatu *cluster* adalah metode koefisien *silhouette*. Evaluasi *clustering* digunakan untuk mengetahui seberapa tepat suatu data dikelompokkan. Tahapan perhitungan koefisien *silhouette* adalah sebagai berikut:

1. Untuk masing-masing objek *i*, hitung rata-rata jarak dari suatu objek ke-*i* dengan semua objek yang berada pada satu kelompok yang sama.
2. Kemudian untuk masing-masing objek *i*, hitung rata-rata jarak suatu objek ke-*i* dengan semua data yang berada pada kelompok yang berbeda, kemudian ambil nilai yang paling kecil.
3. Selanjutnya menghitung nilai dari koefisien *silhouette* berdasarkan Persamaan (16).

$$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)} \tag{16}$$

dengan,

a_i : rata-rata jarak objek ke-*i* dengan semua objek pada satu kelompok yang sama

b_i : rata-rata jarak objek ke-i dengan semua objek pada kelompok yang berbeda
S_i : nilai koefisien *silhouette*.

Hasil perhitungan nilai koefisien *silhouette* dapat bervariasi antara -1 hingga 1. Hasil *cluster* dikatakan tepat jika nilai koefisien *silhouette* adalah 1, berarti objek ke-i sudah berada dalam *cluster* yang tepat. Jika nilai koefisien *silhouette* adalah 0 maka objek ke-i berada di antara dua *cluster* sehingga objek tersebut tidak jelas harus di masukkan ke dalam *cluster* U atau *cluster* V. Jika nilai koefisien *silhouette* adalah -1 artinya struktur *cluster* yang dihasilkan tidak baik, sehingga objek ke-i lebih tepat di masukkan ke dalam *cluster* yang lain [14].

Kriteria subjektif kualitas pengelompokkan berdasarkan nilai *silhouette* yang dibuat oleh Kauffman dan Rousseuw (1990) ditampilkan pada Tabel 1.

Tabel 1. Kriteria subjektif kualitas pengelompokkan berdasarkan koefisien *silhouette*

Nilai *Silhouette*	Interpretasi
0,71-1,00	*Strong Cluster*
0,51-0,70	*Good Cluster*
0,26-0,50	*Weak Cluster*
0,00-0,25	*Bad Cluster*

2.7. Jumlah Penduduk

Jumlah penduduk adalah banyaknya manusia yang bertempat tinggal atau berdomisili pada suatu wilayah atau daerah dan memiliki pekerjaan yang tetap serta terdaftar secara sah berdasarkan peraturan perundang-undangan yang berlaku. Provinsi Kaltim merupakan Provinsi terluas kedua setelah Papua di Indonesia. Provinsi Kaltim memiliki potensi sumber daya alam yang besar, akan tetapi sebagian besar potensi tersebut belum digunakan secara maksimal. Jumlah penduduk di Provinsi Kaltim pada Tahun 2003 sebesar 2.311.162 jiwa, sedangkan pada sensus penduduk Tahun 2010 jumlah penduduk Provinsi Kaltim mencapai 3.047.500 jiwa. Sehingga dalam kurun waktu tersebut jumlah penduduk Provinsi Kaltim meningkat secara signifikan yaitu sebesar 736.338 jiwa, dengan pertumbuhan penduduk setiap tahun rata-ratanya adalah 3,60%. [15].

3. Hasil dan Pembahasan

3.1. Data Penelitian

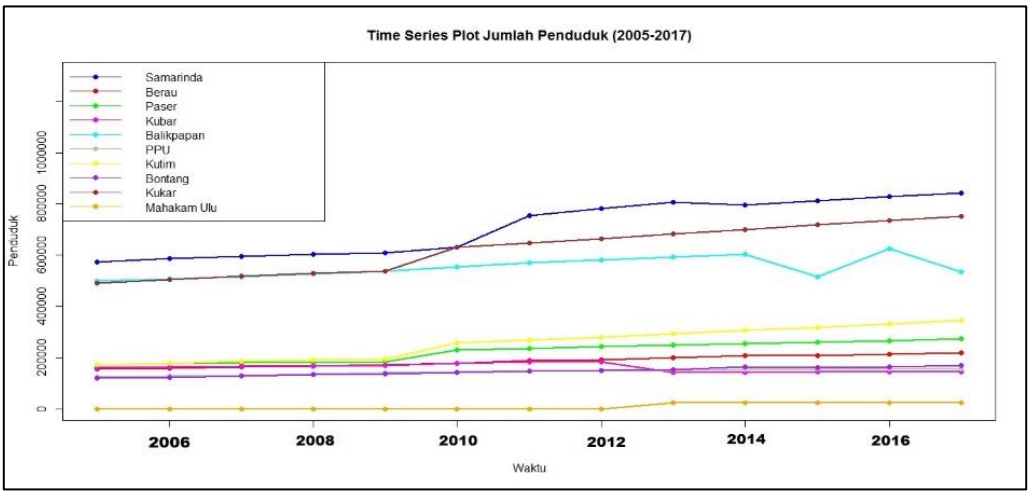
Variabel yang digunakan dalam penelitian ini adalah jumlah penduduk Provinsi Kaltim dalam satuan jiwa pada Tabel 2, yang direkapitulasi oleh Badan Pusat Statistik (BPS) Provinsi Kaltim dan dinotasikan $Z_i(t)$ dengan i=1,2,...,10 [16].

Tabel 2. Data jumlah penduduk Provinsi Kaltim tahun 2005-2017

Tahun	Samarinda	Berau	Paser	Kubar	Balikpapan	PPU	Kutim	Bontang	Kukar	Mahakam Ulu
2005	574439	157453	174420	161582	500406	127586	174018	120348	491607	0
2006	587744	160399	177910	164914	508120	130668	179864	125187	505380	0
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
2017	843446	220601	274206	146998	536012	157711	347468	170611	752091	26089

3.2. Statistika Deskriptif

Pembahasan akan diawali dengan menampilkan statistika deskriptif berupa *time series plot* untuk masing-masing Kabupaten/Kota di Provinsi Kaltim yang ditampilkan pada Gambar 2.



Gambar 2. *Time series plot* jumlah penduduk Provinsi Kaltim

Berdasarkan Gambar 2, diketahui perkembangan jumlah penduduk setiap tahunnya di Kabupaten/Kota Provinsi Kaltim. Kabupaten Mahakam Ulu baru terbentuk di Tahun 2013, yang merupakan hasil pemekaran wilayah Kabupaten Kutai Barat. Jumlah penduduk dihampir semua Kabupaten/Kota mengalami peningkatan setiap tahunnya, salah satu contohnya adalah Kota Samarinda yang merupakan ibukota dari Provinsi Kaltim. Kota Samarinda mengalami perkembangan jumlah penduduk yang pesat bila dibandingkan dengan Kabupaten/Kota lainnya di Provinsi Kaltim.

3.3. Analisis *Cluster*

Metode pengelompokkan yang digunakan adalah pengelompokkan hierarki dengan semua algoritma yang ada pada metode *agglomerative* (pemusatan). Adapun tahapannya sebagai berikut:

1. Normalisasi data

Normalisasi data disini bertujuan untuk membuat semua variabel penelitian berada dalam jangkauan yang sama dan memperkecil perbedaan antar variabel penelitian. Dalam melakukan normalisasi data digunakan Persamaan (1), (2) dan (3) sehingga diperoleh hasil perhitungan yang ditampilkan pada Tabel 3.

Tabel 3. Normalisasi data jumlah penduduk tahun 2005-2017

Tahun	Samarinda	Berau	Paser	Kubar	Balikpapan	PPU	Kutim	Bontang	Kukar	Mahakam Ulu
2005	-1,25	-1,39	-1,31	-0,05	-1,27	-1,75	-1,31	-1,59	-1,39	-0,76
2006	-1,13	-1,26	-1,22	0,16	-1,08	-1,45	-1,21	-1,31	-1,25	-0,76
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
2017	1,24	1,47	1,27	-1,00	-0,39	1,18	1,43	1,37	1,34	1,22

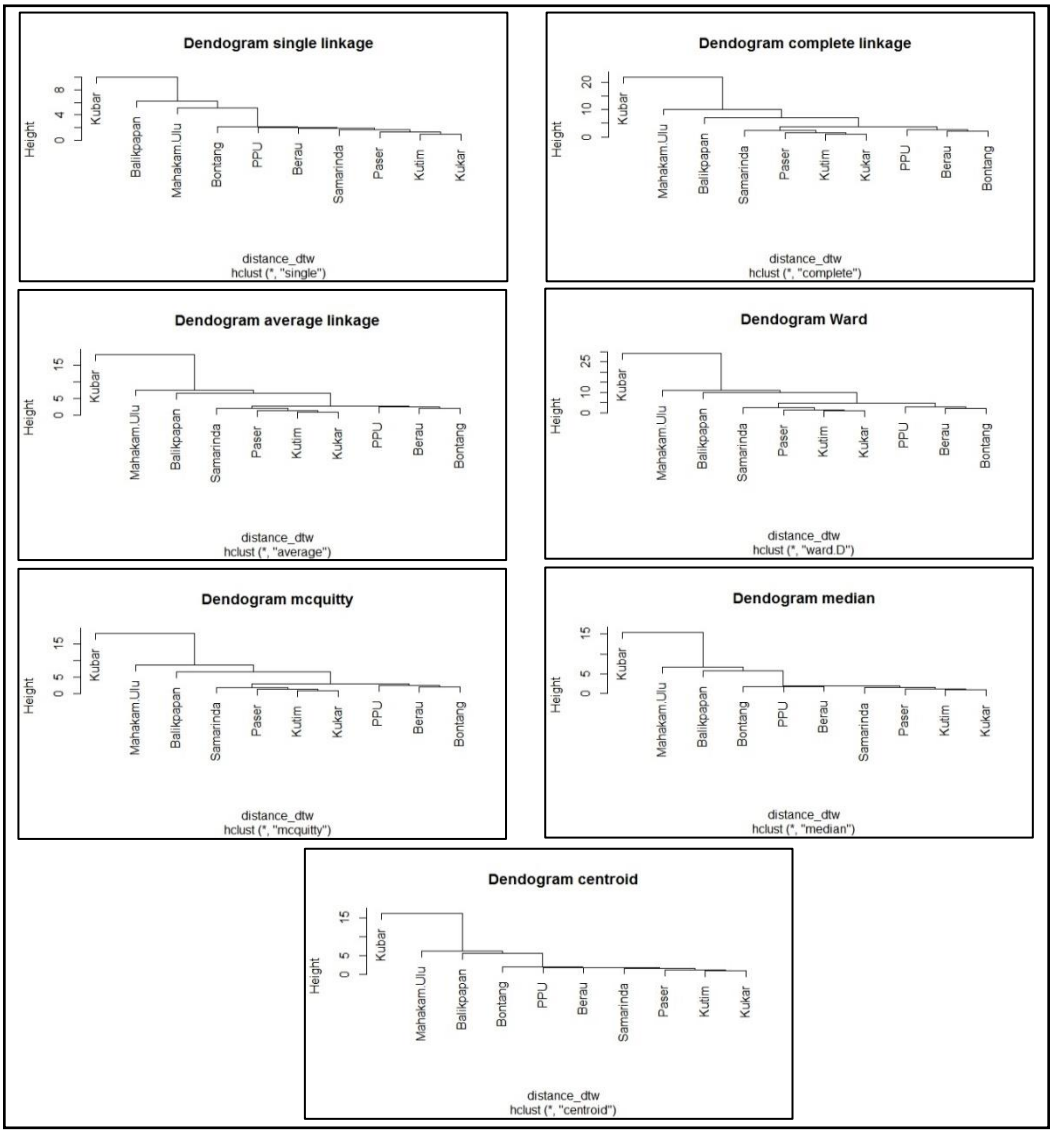
2. Metode *agglomerative* jarak DTW

Jarak pengukuran kemiripan yang digunakan adalah jarak DTW dengan perhitungan menggunakan Persamaan (4), (5) dan (6) dan dilakukan terhadap setiap variabel penelitian, sehingga terbentuk matriks jarak **D** yang ditampilkan pada Tabel 4.

Tabel 4. Hasil perhitungan jarak DTW terhadap setiap variabel penelitian

	Samarinda	Berau	Paser	Kubar	Balikpapan	PPU	Kutim	Bontang	Kukar	Mahakam Ulu
Samarinda	1,00	2,70	1,65	19,33	6,79	3,28	2,37	3,76	2,05	5,07
Berau	2,70	1,00	2,80	19,35	6,13	2,54	2,19	2,15	1,95	6,74
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
Mahakam Ulu	5,07	6,74	8,89	21,95	10,10	7,76	7,93	6,62	7,79	1,00

Langkah selanjutnya setelah melakukan perhitungan jarak DTW adalah melakukan proses pengelompokkan menggunakan semua algoritma pada metode *agglomerative*, sehingga diperoleh dendogram yang ditampilkan pada Gambar 3.



Gambar 3. Dendogram metode *agglomerative* jarak DTW

3. Metode *agglomerative* jarak ACF

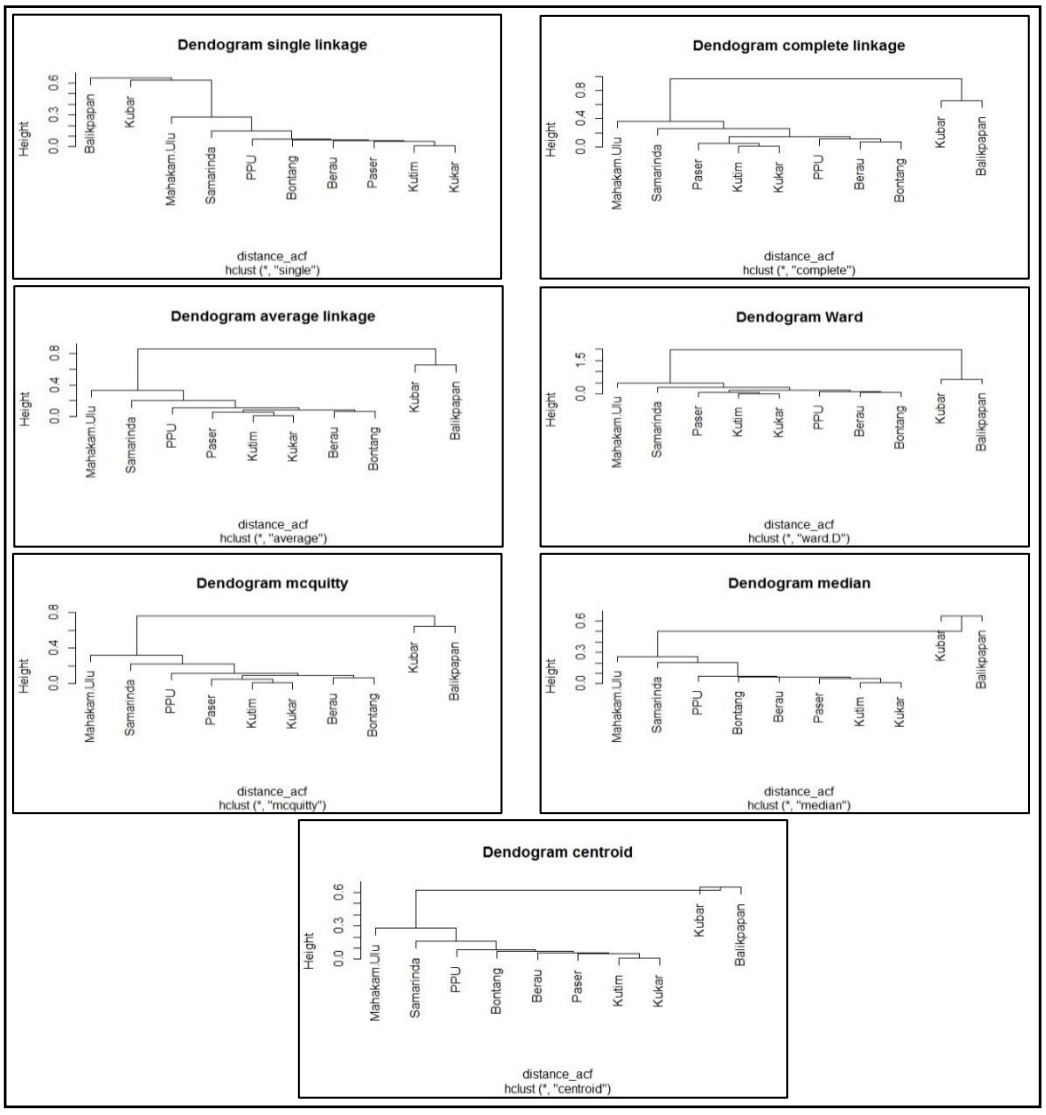
Jarak pengukuran kemiripan yang digunakan adalah jarak ACF dengan perhitungan menggunakan Persamaan (7).

Hal ini dilakukan terhadap setiap variabel penelitian, sehingga terbentuk matriks jarak **D** yang ditampilkan pada Tabel 5.

Tabel 5. Hasil perhitungan jarak ACF terhadap setiap variabel penelitian

	Samarinda	Berau	Paser	Kubar	Balikpapan	PPU	Kutim	Bontang	Kukar	Mahakam Ulu
Samarinda	1,00	0,19	0,15	0,91	0,84	0,26	0,18	0,24	0,18	0,31
Berau	0,19	1,00	0,10	0,96	0,87	0,12	0,06	0,07	0,06	0,36
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
Mahakam Ulu	0,31	0,36	0,28	0,63	0,67	0,33	0,33	0,36	0,32	1,00

Langkah selanjutnya setelah melakukan perhitungan jarak ACF adalah melakukan proses pengelompokkan menggunakan semua algoritma pada metode *agglomerative*, sehingga diperoleh dendogram yang ditampilkan pada Gambar 4.



Gambar 4. Dendogram metode *agglomerative* jarak ACF

4. Penentuan jarak pengukuran kemiripan terbaik dan algoritma yang optimal

Pengujian validitas pada penelitian ini bertujuan untuk menghasilkan proses *clustering* yang optimal, artinya proses pembentukan *cluster* dengan algoritma pengelompokkan yang bermacam-macam didasarkan pada jarak pengukuran kemiripan serta jumlah kelompok (*cluster*) yang optimal. Uji validitas jarak dan algoritma yang digunakan dalam penelitian ini adalah koefisien korelasi *cophenetic* berdasarkan Persamaan (15) dan ditampilkan pada Tabel 6.

Tabel 6. Perbandingan setiap algoritma pengelompokkan dan jarak pengukuran kemiripan

Jarak	Algoritma						
	Single	*Complete*	*Average*	*Ward*	*WPGMA*	*Median*	*Centroid*
DTW	0,94	0,96	0,97	0,96	0,97	**0,97**	0,97
ACF	**0,97**	**0,98**	**0,99**	**0,97**	**0,98**	0,95	**0,98**

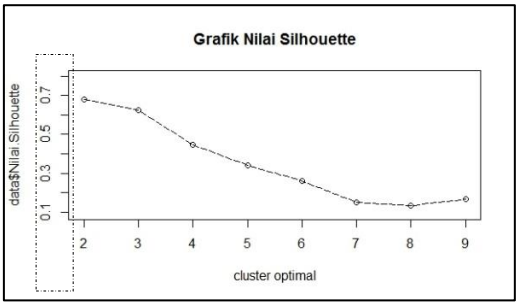
Berdasarkan Tabel 6, diketahui bahwa nilai koefisien korelasi *cophenetic* untuk setiap penggunaan jarak yang berbeda terhadap setiap algoritma pengelompokkan yang ada didalam metode pengelompokkan hierarki (*agglomerative*) sangat beragam. Nilai dari koefisien korelasi *cophenetic* berkisar antara -1 sampai dengan 1, yang artinya ketika nilai koefisien korelasi *cophenetic* mendekati 1 berarti jarak pengukuran kemiripan yang digunakan dalam proses *clustering* cukup baik.

Secara garis besar, dapat disimpulkan bahwa dalam proses pengelompokkan Kabupaten/Kota di Provinsi Kaltim berdasarkan jumlah penduduk, jarak pengukuran kemiripan yang terbaik adalah *autocorrelation based distance* (ACF), dimana jarak ACF menunjukkan nilai yang optimal (hampir mendekati 1) di beberapa algoritma pengelompokkan. Jarak ACF ini nantinya akan digunakan dalam proses analisis selanjutnya untuk menentukan jumlah *cluster* yang representatif dalam proses *clustering*.

Adapun algoritma pengelompokkan yang optimal pada proses pengelompokkan Kabupaten/Kota di Provinsi Kaltim berdasarkan jumlah penduduk adalah algoritma *average linkage*, dikarenakan memiliki nilai koefisien korelasi *cophenetic* yang terbesar diantara algoritma pengelompokkan lainnya yaitu sebesar 0,99.

5. Jumlah *cluster*

Setelah mendapatkan jarak pengukuran kemiripan terbaik yaitu jarak *autocorrelation based distance* (ACF) serta algoritma pengelompokkan yang optimal yaitu *average linkage*, langkah selanjutnya adalah menentukan jumlah *cluster* yang representatif dalam proses *clustering* algoritma *average linkage*. Uji validitas *cluster* yang digunakan pada penelitian ini adalah metode koefisien *silhouette* berdasarkan Persamaan (16). Nilai koefisien *silhouette* ditunjukkan pada Gambar 5.

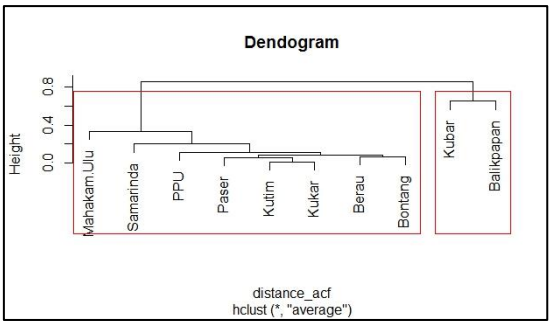


Gambar 5. Grafik nilai *silhouette*

Nilai koefisien *silhouette* dapat bervariasi antara -1 hingga 1. Jumlah *cluster* dikatakan representatif jika nilai koefisien *silhouette* mendekati 1. Berdasarkan Gambar 5, dapat diketahui bahwa jumlah *cluster* yang representatif dalam mengelompokkan Kabupaten/Kota di Provinsi Kaltim adalah 2 *cluster* dengan nilai koefisien *silhouette* terbesar yaitu 0,68. Berdasarkan Tabel 1, nilai koefisien *silhouette* yang diperoleh termasuk kategori *good cluster*.

6. Profilisasi dan interpretasi hasil *cluster*

Langkah selanjutnya adalah melakukan profilisasi dan interpretasi hasil *cluster*. Pada proses pengelompokkan Kabupaten/Kota di Provinsi Kaltim berdasarkan jumlah penduduk menggunakan data runtun waktu, jarak pengukuran kemiripan yang digunakan adalah *autocorrelation based distance* (ACF) serta algoritma pengelompokkan yang optimal yaitu *average linkage* dengan jumlah *cluster* yaitu 2, sehingga diperoleh dendogram pengelompokkan yang dapat dilihat pada Gambar 6.



Gambar 6. Dendogram pengelompokkan Kabupaten/Kota Provinsi Kaltim

Berdasarkan Gambar 6, diketahui hasil pengelompokkan Kabupaten/Kota di Provinsi Kaltim berdasarkan jumlah penduduk menggunakan data runtun waktu. Pada *cluster* 1 terdapat 8 Kabupaten/Kota yang bergabung. Kutai Timur dan Kutai Kartanegara yang pertama kali bergabung menjadi 1 *cluster* dan diikuti oleh Kabupaten/Kota lainnya, yaitu Paser, Berau, Bontang, PPU, Samarinda dan Mahakam Ulu. Sedangkan pada *cluster* 2 terdapat 2 Kabupaten/Kota yang bergabung yaitu Kutai Barat dan Balikpapan, sehingga hasil pengelompokkan bisa diilustrasikan seperti pada Gambar 7.

Gambar 7. Peta Provinsi Kaltim berdasarkan *cluster*

4. Kesimpulan

Pengukuran kemiripan terbaik dalam proses pengelompokkan Kabupaten/Kota di Provinsi Kaltim adalah jarak *autocorrelation based distance* (ACF) dengan algoritma pengelompokkan yang optimal adalah algoritma *average linkage*. Nilai koefisien korelasi *cophenetic* yang diperoleh sebesar 0,99. Jumlah *cluster* yang representatif adalah 2 *cluster* dengan nilai koefisien *silhouette* yang diperoleh sebesar 0,68.

Referensi

[1] Haryati, S., Sudarsono, A., & Suryana, E., 2015, Implementasi Data Mining untuk Memprediksi Masa Studi Menggunakan Algoritma C4.5, Jurnal Media Infotama 11 (2): 130-138
[2] Prasetyo, E., 2012, Data Mining: Konsep dan Aplikasi Menggunakan MATLAB, Yogyakarta: Penerbit Andi
[3] Liao, T. W., 2005, Clustering of Time Series Data Survey, Pattern Recognition 38: 1857-1874
[4] Virgiawan, D.M. & Mukhlash, I., 2013, Aplikasi Association Rule Data Mining untuk Menemukan Pola Data Nilai Mahasiswa Matematika ITS, Jurnal Sains dan Seni ITS 1 (1): 1-6
[5] Mabrur, A.G. & Lubis, R., 2012, Penerapan Data Mining untuk Memprediksi Kriteria Nasabah Kredit, Jurnal Komputer dan Informatika 1 (1): 53-57
[6] Santosa, B., 2007, Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis, Teori dan Aplikasi, Yogyakarta: Graha Ilmu
[7] Supranto, 2010, Statistik: Teori dan Aplikasi Edisi 8, Jakarta: Erlangga
[8] Sartono, B., Affendi, F. M., Sumertajaya, I. M. & Angraeni, Y, 2003, Analisis Peubah Ganda, Bogor: Fakultas Matematika dan Ilmu Pengetahuan Alam IPB
[9] Montero, P. & Vilar, J.A., 2014, TSclust: An R Package for Time Series Clustering, Journal of Statistical Software 62 (1): 01-43
[10] Riyadi, M.A.A., Fithriasari, K. & Dwiatmono, 2016, Data Mining Peramalan Konsumsi Listrik dengan Pendekatan Cluster Time Series sebagai Preprocessing, Jurnal Sains dan Seni ITS 5 (1): 121-126
[11] Johnson, R. A. & Wichern, D.W., 2002, Applied Multivariate Statistical Analysis, Fifth Edition, New Jersey: Pearson Prentice Inc
[12] Minitab Methods and Formulas, (Mei 12, 2019), Citing Internet sources URL https://support.minitab.com
[13] Saracli, S., Dogan, N. & Dogan, I., 2013, Comparison of Hierarchical Cluster Analysis Methods by Cophenetic Correlation, Journal of Inequalities and Applications, doi: 10.1186/1029-242X-2013-203
[14] Kaufman, L. & Rousseeuw, P.J., 1990, Finding Groups in Data An Introduction to Cluster Analysis, New Jersey: John Wiley & Sons Inc Publication
[15] BAPPEDA Provinsi Kalimantan Timur, (Mei 12, 2019), Citing Internet sources URL www.bappedakaltim.com
[16] BPS Provinsi Kalimantan Timur, (Mei 05, 2019), Citing Internet sources URL www.kaltim.bps.go.id